图书情报工作
第63卷 第9期 2019年5月

临床医学课程知识主题图谱构建研究*

■ 陆泉[1,2] 谢祎玉[1] 陈静[3] 张涵[1] 崔浩冉[1] 聂书源[1]

[1] 武汉大学信息资源研究中心 武汉 430072 [2] 武汉大学大数据研究院 武汉 430072

[3] 华中师范大学信息管理学院 武汉 430079

摘要：［目的/意义］从知识主题的角度切入，建立全面的课程知识体系，解决现有课程体系设计和教学中的课程间知识点重复及“知识孤岛”问题，从而有效开展专业知识服务。［方法/过程］以临床医学专业主干课程为研究对象，基于医学主题词表、电子教材、电子教案等医学教育数据，通过LDA模型挖掘课程中的知识主题，利用关联分析揭示课程间、知识主题间及课程与知识主题间的细粒度关联，从而构建临床医学课程知识主题图谱。［结果/结论］研究从专业课程体系与知识主题视角构建出领域知识图谱，有助于教学管理人员及师生掌握专业知识体系，开展知识导向型教学活动，推进医学领域知识组织与服务及智慧医学教育发展。

关键词：课程知识主题图谱 知识图谱 LDA 关联分析 临床医学

分类号：G251

DOI：10.13266/j.issn.0252-3116.2019.09.011

在医学类课程的学习过程中，往往涉及许多知识互通的现象，课程与知识点之间的学习存在一定的层次顺序，在学习新知识时需要结合以往学过的知识作为补充。但医学类学科现有的课程体系设计存在课程和知识点重复[1]、课程学习中存在有关知识主题的“知识孤岛”等问题，教师难以有效组织专业知识体系教学活动，学生无法快速定位与现学知识点相关的内容。因此，如何避免知识点的重复、聚焦关键核心知识点，建立起整个学科专业的体系知识结构，是医学专业相关人员的一大需求，也是科学设置课程、优化课程体系、深化知识组织与服务亟需解决的重要基础教育技术问题。

临床医学专业是一门实践性很强的应用科学专业，致力于培养具备基础医学、临床医学的基本理论和医疗预防的基本技能，能在医疗卫生单位、医学科研等部门从事医疗及预防、医学科研等方面工作的医学高级专门人才[2]。临床医学专业课程包括解剖学、生理学、内科学、外科学等多门课程，知识点多，体系庞大，因此需要建立全面的课程知识体系，从而有效开展专业知识服务。

1 研究现状

课程体系建设是促进教育改革与发展的重要抓手[3]，因此，对课程体系的研究一直是教学改革的热门所在。对于课程体系中的知识点重复及“知识孤岛”问题，许多学者探讨了其解决方法。胡文韬[4]基于知识图谱，对学生从学习目标开始到学习路径构建过程中的课程选择和课程排序进行研究，试图建立课程间的联系，发现课程的知识结构，从而解决学生在学习过程中的信息过载和知识迷航问题。叶春森等[5]依据知识管理理论，提出基于知识地图的知识集成模式，为降低知识内耗、控制知识集成过程、消除“知识孤岛”提供了新方法。郑宁[6]基于自然语言处理技术获取算法知识名称并构建本体来识别网络程序资源中的算法知识点，从而将海量网络程序资源按知识结构组织起来，解决其中存在的“知识孤岛”现象。在课程体系架构及建设方面，商玮等[7]借鉴基于工作过程的课程开发思路与CDIO工程教育模式，在融入教学工厂理念的

* 本文系教育部人文社会科学重点研究基地重大项目“大数据资源的挖掘与服务研究——面向医疗健康领域”（项目编号：17JJD870002）研究成果之一。

作者简介：陆泉（ORCID：0000-0002-8679-9866），教授，博士；谢祎玉（ORCID：0000-0003-1956-1038），硕士研究生；陈静（ORCID：0000-0002-6444-2962），副教授，博士，通讯作者，E-mail：jchen@mail.ccnu.edu.cn；张涵（ORCID：0000-0003-2152-5378），本科生；崔浩冉（ORCID：0000-0001-8914-3443），本科生；聂书源（ORCID：0000-0002-7112-6773），本科生。

收稿日期：2018-09-12 修回日期：2018-11-27 本文起止页码：101-108 本文责任编辑：刘远颖

基础上构建了 TF-CDIO 电子商务专业课程体系，周明等[8]研究了大数据视角下信息管理专业课程体系的创新建设，提出从大数据发展的角度着手寻找专业特色，构建新的课程体系。就临床医学专业而言，李莉等[9]分析了医疗大数据的价值与教学之间的关系，认为临床大数据的应用将改变传统的眼科临床教学体系。

然而，课程是知识主题的组织形式，知识主题是课程的核心内容，课程体系的建设与利用必须建立在对专业知识体系的深度挖掘与全盘掌握基础之上。R. J. Todd [10]研究了学生如何利用现有的课程知识主题将发现的信息转为个人知识，并绘制和衡量学生对课程主题知识的变化。朱珂等[11]使用主题图技术对单个网络课程知识组织方式进行重组，对知识点进行多粒度、多层次的组织，实现网络课程知识点语义关联和智能分类，为个性化学习等学习模式提供支持。E. Melis 等[12]开发了一种基于网络的通用学习系统 ActiveMath 来为每个知识主题构建学习资料，即由学习者选择目标知识主题，系统为知识主题选择相关资料，从而为学习者生成整个课程。

综上所述，虽然有部分学者在研究课程体系时研究了知识主题的获取与表达，但尚未见从知识主题切入，将课程体系与知识主题形成映射图谱，并对课程与知识主题之间的定量关系进行研究，因此本研究对临床医学教育数据进行深度挖掘，利用 LDA 模型挖掘课程中的知识主题，关联分析法揭示课程间、知识主题间及课程与知识主题间的细粒度关联，从专业课程体系与知识主题视角来研究与构建临床医学课程知识主题图谱，有助于教学管理人员及师生掌握专业知识体系，开展知识导向型教学活动，推进医学领域知识组织与服务及智慧医学教育发展。

2 临床医学课程知识主题图谱模型构建

临床医学课程知识主题图谱模型主要包括 3 个子模块：①临床医学教育数据预处理。主要对研究所需数据进行收集、分词及去停用词等操作，从而得到模型的输入文件。②LDA 主题挖掘。利用 LDA 算法挖掘出文本中的知识主题。③关联计算。结合挖掘到的知识主题，计算主题词间关联及章节与知识主题间的关联度权重等。

2.1 临床医学教育数据预处理

预处理过程是针对临床医学课程原始文本进行加工，如医学主题词表、电子教材、电子教案等医学教育数据，最终生成 LDA 主题挖掘所需要的数据格式。临床医学教育数据预处理模块的具体流程如图 1 所示：

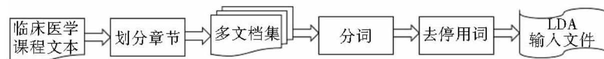


图 1 临床医学教育数据预处理流程

2.1.1 数据来源与采集 本研究通过调研武汉大学临床医学（五年制）本科人才培养方案及其课程体系[13]，选取该专业 14 门主干课程作为研究对象，课程包括解剖学、组织胚胎学、生理学、生物化学与分子生物学、药理学、病理学、病理生理学、医学微生物学、医学免疫学、临床技能学、内科学、外科学、妇产科学和儿科学，收集其课程简介、电子教材、电子教案、课程大纲等课程资料，依据人民卫生出版社第八版教材的目录对课程章节进行划分，并将对应的课程资料转换为文本格式，共获得 385 个课程章节文本。

2.1.2 分词及去停用词 知识主题词是本研究的基本单元，因此需要对文本进行分词以得到 LDA 算法的输入文件。使用 Python 爬取中国生物医学文献数据库[14]中主题检索的主题词作为分词字典，整合“哈工大停用词词库”“百度停用词表”等停用词表，去重后得到一份较为全面的停用词表，采用开源中文分词工具 jieba 进行分词，并将分词后文本按照之前同样的方式进行划分，得到 385 个分词后的课程章节文本，每个章节文本即为 LDA 主题挖掘模块的输入文档。

2.2 LDA 主题挖掘

LDA（Latent Dirichlet Allocation）是 D. M. Blei 等[15-16]提出的一种文档主题生成模型，包含词、主题和文档三层结构，可以用来识别文档集中的潜在主题信息。LDA 采用词袋（bag of words）方法，将每篇文档看作一个词频向量，文档是由若干个主题混合组成，每个主题是一个关于词的概率分布。对于给定的文档集 $D=\{d_1,d_2,\cdots,d_n\}$，由给定的先验 Dirichlet 分布，得到文档生成的似然函数，其过程如下[17]：

（1）对 D 中的每个文档 d，由 $\theta_d \sim Dirichlet(\alpha)$，得到文档 d 上主题的多项式向量 θ_d。

（2）对每个主题 z，由 $\varphi_z \sim Dirichlet(\beta)$，得到主题 z 上的词汇的多项式向量 φ_z。

（3）对文档 d 中的词汇 $w_{d,i}$，生成一个主题 z_j 服从参数为 θ_d 的多项式分布，根据特定的主题比例 β，生成词汇 $w_{d,i}$ 的概率分布 $P(w_{d,i}|z_j,\beta)$。

对文档集 D，LDA 主题抽取过程可以总结为根据 θ 和 z，求出使 $P(D|\alpha,\beta)$ 最大的参数 α 和 β，其中：

$P(D \mid \alpha,\beta) = P(\theta,z \mid \alpha,\beta) = P(\theta \mid \alpha) \prod_{i=1}^{N} P(z_i \mid \theta) P(w_i \mid z_i,\beta)$　　公式(1)

采用 Gibbs 抽样[18]对上述公式中的隐含变量进行参数推断,从而计算后验概率。Gibbs 抽样过程中得到主题—主题词和文档—主题两个矩阵,将其与 $P(z_i \mid z_i,w_i)$ 循环迭代计算,当数值收敛时的分布即主题的对应分布。根据 θ 和 z 则可得到文档中每个主题的概率分布及主题中每个主题词的概率分布。通过概率计算则可得到每个文档中的知识主题词。

LDA 主题挖掘模块基于 LDA 算法对预处理得到的文本进行主题挖掘,能够利用文档的潜在语义信息得到知识主题词。朱泽德等[19]的研究也表明基于 LDA 的关键词抽取方法能够较好地避免将常用词作为关键词,并解决词未能全面准确覆盖文档主题信息的问题,提高关键词抽取的准确率。LDA 主题挖掘模块的具体过程如图 2 所示:

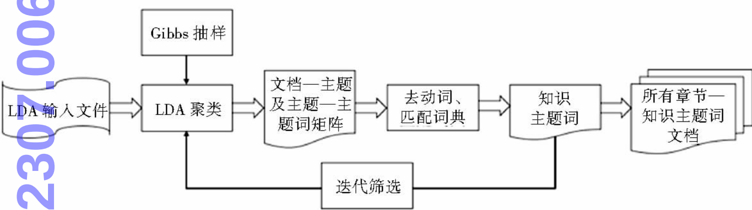


图 2　LDA 主题挖掘流程

本研究的主要目的是得到文档中的知识主题词,不需要进行主题分类,因此参考唐晓波等[20]对微博热点挖掘的参数设置,每篇文档提取出的主题数为 $k = 10$。根据文献调研[21-22]及经验值确定 α 和 β 的取值,设定 $\alpha = 50/k$,$\beta = 0.01$。Gibbs 循环迭代抽样的最大次数设为 1 000 次。实验结果表明以上参数设置在文档集中有较好的表现。然后,根据得到的主题—主题词和文档—主题两个矩阵,对主题词进行筛选,筛选规则为:对于文档集 $D = \{d_1, d_2, \cdots, d_{385}\}$ 中的每个文档 d,其中的主题 z_j 权重为 v_j,z_j 中主题词 w_i 权重为 v_i,则主题词 w_i 在文档 d 中的最终权重为 $v_i \times v_j$,对每个词的权重进行排序,将权重大于等于设置阈值的词作为文档的主题词,在本实验中设置权重阈值为 0.008 能够达到较理想的效果。

按照上述参数设置及规则进行一次计算后,得到每个文档中符合条件的知识主题词,共有 385 个章节—知识主题词文档,每个文档记录了对应章节的知识主题词。但由于 LDA 算法生成主题词是一个随机过程,每次计算得到的知识主题词有细微差异,因此需要进行多次迭代实验,观察得到的主题词效果。迭代方法如下:

(1)对 D 中的文档 d,首次计算得到的知识主题词集合为 w_1,按照规则再次计算得到的知识主题词集合为 w_2,更新 d 的知识主题词为两者的并集,即 $w_1 = w_1 \cup w_2$。

(2)重复第一步的计算 n 次,直到每个文档的 w_1 中不加入新词,达到稳定状态。

本研究经过 7 次迭代实验,每个文档的知识主题词集合达到稳定状态。然后,随机抽取 50 个文本,对照该章节的教学大纲观察得到的知识主题词,发现挖掘得到的结果能够作为该章节主题的概括。

在得到每个文档的知识主题词后,考虑到 LDA 算法会挖掘出一些非知识主题词,且大部分知识主题词应为名词,因此为确保主题词的可用性及可靠性,本研究结合《现代汉语动词表》对提取出的知识主题词进行去动词处理,改进 LDA 主题挖掘的效果。最终得到每个章节的知识主题词,形成临床医学课程知识主题的章节—知识主题词多对多映射矩阵,揭示出章节中包含的知识主题及知识主题覆盖的章节,最终得到 1 696 个不重复的知识主题词。表 1 显示了部分章节的知识主题词信息。

表 1　部分章节知识主题词

章节编号	章节名称	所属课程	知识主题词
1	疾病概论	病理生理学	亚健康、发病学、神经机制、体液机制、病因学、先天性、免疫性、转归
50	循环系统疾病	儿科学	儿科、血液、循环系统、心脏、先天性心脏病、动脉、血管、心房、静脉
100	盆部与会阴	解剖学	盆部、会阴、骨盆、盆壁肌、盆腔脏器、盆筋膜、筋膜间隙、肛管
150	能量代谢与体温	生理学	能量、代谢、体温、血糖、缺氧、肌肉、蛋白质、脂肪、产热
200	颅内和椎管内肿瘤	外科学	外科、颅内肿瘤、椎管内肿瘤、胶质瘤、脑膜瘤、听神经瘤、垂体瘤、淋巴瘤

2.3　关联计算

在得到课程知识主题的多对多映射矩阵的基础上,对 N 个知识主题词统计共现的文本频数,形成知识主题词的 $N \times N$ 共现矩阵,根据共现矩阵进行关联分析,从而揭示课程间细粒度知识主题关联。关联计算模块主要对主题—主题关联、主题—章节关联和章

节—章节关联三个方面进行计算。

2.3.1 主题—主题关联计算 关联分析是知识发现的一种手段,可以量化地描述物品 A 的出现对物品 B 的出现有多大的影响[23],通常用于事务数据库如销售数据中。将关联分析应用于医学领域,可以从繁杂的医学资料中挖掘出有价值的信息。张晗等[24]应用关联规则算法分析抗肿瘤药物主题词和副主题词组配模式,抽取出主题词的依存关系及五类药物相关的语义关系组合。如某篇关于药物治疗的文献标引中,包含"病 A/药物治疗"主题词的同时也存在"药 B/治疗应用"主题词,则表明药 B 可能具有治疗病 A 的功效。因此对于本研究的课程知识主题数据,一个课程章节可以看作是一个事务 T,由多个知识主题词的项集组成。为得到知识主题词之间的语义关联,可以对其共现矩阵进行关联分析,挖掘出满足一定支持度和可信度条件下的频繁出现在一起的知识主题词[25]。

主题—主题关联计算基于 Apriori[26] 算法。对 LDA 主题挖掘模块得到的 385 个知识主题词文本及 1 696个知识主题词:首先统计每个知识主题词出现的文本数,如得到 A 词和 B 词出现的文本数分别为 C_A 和 C_B,再根据共现矩阵得到每个词对{A,B}在所有文本中共现的文本总数 $C_{A\cap B}$。对每个有向词对{$A\rightarrow B$},得到支持度大于等于最小支持度,可信度大于等于最小可信度,同时作用度大于 1 的关联规则[27-28]。

支持度描述词 A 和词 B 在所有文本中同时出现的概率,计算公式为:

$$Support(A\rightarrow B) = P(A\cap B) = C_{A\cap B}/385 \qquad 公式(2)$$

可信度描述出现词 A 的文本,同时也出现词 B 的概率,计算公式为:

$$Confidence(A\rightarrow B) = P(B|A) = C_{A\cap B}/C_A \qquad 公式(3)$$

作用度描述词 A 对词 B 的影响程度,作用度大于 1 则是正相关,计算公式为:

$$Lift(A\rightarrow B) = \frac{P(B|A)}{P(B)} = \frac{C_{A\cap B}\cdot 385}{C_A\cdot C_B} \qquad 公式(4)$$

在本研究中,最小支持度取 0.002,最小可信度取 0.5。根据上述算法计算,共得到 12 055 条强关联规则,描述了一个知识主题词对另一个知识主题词的单向关联度。在本研究中定义主题—主题之间的关联类型有三种,即基础关系、进阶关系和同级关系,根据关联计算的结果,在得出的所有关联规则中,有以下 3 种情况:

(1)若词对{A,B}只存在一条关联规则,即 $A\rightarrow B$,可信度为 x,说明主题词 A 影响主题词 B 的出现,因此定义 A 为 B 的基础主题,B 为 A 的进阶主题,即在学习主题词 B 之前需要先具备主题词 A 的知识,学习主题词 A 之后可以去继续学习主题词 B 的知识。

(2)若词对{A,B}存在两条关联规则,即存在 $A\rightarrow B$,可信度为 x,又存在 $B\rightarrow A$,可信度为 y,且 $x>y$,则说明主题词 A 对主题词 B 出现的影响大于主题词 B 对主题词 A 出现的影响,因此舍弃 B→A 这条规则,定义 A 为 B 的基础主题,B 为 A 的进阶主题;反之若 $x<y$,则 B 为 A 的基础主题。

(3)若词对{A,B}存在两条关联规则,即存在 $A\rightarrow B$,可信度为 x,又存在 $B\rightarrow A$,可信度为 y,且 $x=y$,则说明主题词 A 与主题词 B 具有同等影响,因此合并两条规则,定义 A 和 B 为同级主题,即主题词 A 和主题词 B 可以并行学习。

根据上述规则和主题间的三种关系,对 12 055 条关联规则进行删除及合并后,得到 8 933 条有效关联,其中同级关系 6 632 条,基础关系和进阶关系 2 301 条。如表 2 显示了"病因学"与其他知识主题词的关系。

表 2 "病因学"与其他知识主题词的关系

关系	知识主题词
进阶关系	发病学、免疫
基础关系	分子机制、死亡、体液机制、神经机制、组织细胞机制先天性、脑死亡、亚健康
同级关系	转归、症状

2.3.2 主题—章节关联计算 主题—章节关联揭示章节中各个主题所占权重大小,即知识主题词的重要程度,计算基于 TF-IDF 算法[29]。对 385 个包含知识主题词文本,首先计算每个主题词相对于 385 个章节的 IDF(逆文本频率指数)值,对于主题词 i,包含 i 的文本总数为 df_i,则:$IDF(i) = log(385/df_i)$;然后,计算主题词 i 在经过预处理模块后的对应章节文本 j 中的 TF(词频)值,并根据本研究的特点进行归一化处理,即 $TF(i,j) = N_{i,j}/N_j$,其中 $N_{i,j}$ 为主题词 i 在文本 j 中出现的次数,N_j 为文本 j 中的总词数;则主题词 i 在文本 j 中的 TF-IDF 值为:

$$TF-IDF(i,j) = TF(i,j)\cdot IDF(i) \qquad 公式(5)$$

此外,本研究还计算了主题—课程关联,主题—课程关联揭示课程中知识主题词的重要程度。将 385 个经过预处理模块后章节文本按照课程合并,划分为 14 个文本,将 385 个经过 LDA 主题挖掘模块后的章节主题词文本按照课程合并,并去除重复主题词,得到 14 门课程中的知识主题词文本。主题—课程关联与主

题—章节关联计算方法类似，计算每个主题词相对于 14 门课程的 IDF 值和 TF 值，并进行 TF-IDF 的计算与归一化。

2.3.3　章节—课程关联计算　章节—课程关联揭示课程中各章节的重要程度。从逻辑层面上将，章节的知识主题词可以看作对章节内容的高度凝练，因此章节—课程关联计算建立在主题—课程关联计算的基础上。对于每一章节，计算其包含的所有主题词在所属课程中的 TF-IDF 值之和，将求和结果除以章节包含的主题词数进行平均化处理。对于课程 c 的章节 j，章节 j 中包含主题词个数为 n，则章节 j 在课程 c 中的权重为：

$$w(j,c)=\frac{\sum_{i=0}^{n} TF-IDF(i,c)}{n}$$ 公式(6)

在得到课程 c 的所有章节权重后，对章节 j 的权重进行归一化处理，归一化方式为章节的 j 权重值除以该课程全部章节的权重值之和，课程 c 中包含 k 个章节，则章节 j 在课程 c 中的最终权重为：

$$weight(j,c)=\frac{w(j,c)}{\sum_{i=0}^{k} w(k,c)}$$ 公式(7)

2.3.4　章节—章节关联计算　由于知识主题词是对应章节的高度凝练，因此章节—章节关联计算建立在主题—主题关联计算的基础上。通过计算两个章节间所有主题词的关联度之和，并根据两个章节间可能存在的关联规则数将其平均化，即可用来表示对应的章节—章节关联。假设章节 A 有 x 个知识主题词，章节 B 有 y 个知识主题词，两个章节的知识主题词中共出现 z 条关联规则，每条规则的可信度为 c，提出章节 A 与章节 B 的关联权重计算公式为：

$$w(A,B)=\frac{\sum_{i=0}^{z} c_i}{x\cdot y}$$ 公式(8)

由上述公式计算得到所有章节—章节关联度，并按照关联度由高到底的顺序格式化数据，总结归纳出每个章节的关联章节。

3　研究结果

通过对武汉大学临床医学五年制的培养方案和课程体系的调研，本研究选取的 14 门专业主干课程可以分为基础医学课程、过渡课程和临床医学课程三类，其具体包含的课程如表 3 所示。3 种类型课程存在偏序有向性，课程之间的学习具有一定的逻辑和时间顺序，临床医学课程必须建立在基础医学课程和过渡课程已学习过的基础上。

表 3　3 种类型课程

类型	课程
基础医学课程	解剖学、组织胚胎学、生理学、生物化学与分子生物学、药理学、病理学、病理生理学、医学微生物学、医学免疫学
基础医学、临床医学过渡课程	临床技能学
临床医学课程	内科学、外科学、妇产科学、儿科学

本研究旨在通过对课程资料中知识主题的挖掘和分析，来构建临床医学专业的课程知识主题图谱，从而辅助师生直观了解重要知识点，建立全面的课程知识体系，提高教学质量和学习效果。在挖掘得到临床医学中的知识主题词及课程—章节—主题三者间关联后，得到主题词 1 696 个，主题—主题关联 8 933 条，章节—主题关联 4 194 条，主题—课程关联 3 308 条，章节—课程关联 385 条，章节—章节关联 16 120 条，以数据库文件形式构建临床医学课程知识主题知识库，用于存储研究中涉及的所有数据。

在得到临床医学课程知识主题知识库的基础上，本研究采用力导向图来实现课程知识主题图谱结构的可视化呈现[30]，并利用百度开源工具 Echarts 完成力导向图的创建。下面从临床医学课程知识主题图谱总览、章节—章节关联、主题—主题关联 3 个方面描述研究结果。

3.1　临床医学课程知识主题图谱总览

临床医学课程知识主题图谱总体以“临床医学五年制”节点为中心向外辐射为三层，如图 3 所示，从内到外第一层较大的节点表示课程，第二层节点表示章节，第三层叶子节点表示知识主题，其中连线代表课程—章节—主题三者之间的关联。

在图 3 中仅显示权重较大、关联较为密切的节点，能够清晰直观地聚焦核心知识点。如在《内科学》课程中，“泌尿系统疾病”“呼吸系统疾病”“血液系统疾病”和“心内科”4 个章节为该课程的重点章节，在“心内科”章节中的重要知识主题词有“动脉”“心肌”“心脏”和“心室”，同时可以看到该章节与《病理生理学》的“心力衰竭”章节及《临床技能学》中的“心电图学”章节有较为密切的关联。根据课程之间的偏序有向性，在学习“心内科”之前需要具备“心力衰竭”和“心电图学”章节的相关知识点，其中主题词“心肌”是章节“心内科”和“心力衰竭”共有的主题词，“心电图”是章节“心内科”和“心电图学”共有的主题词。

3.2　章节—章节关联

对于课程中的一个章节节点，图谱呈现与该章节

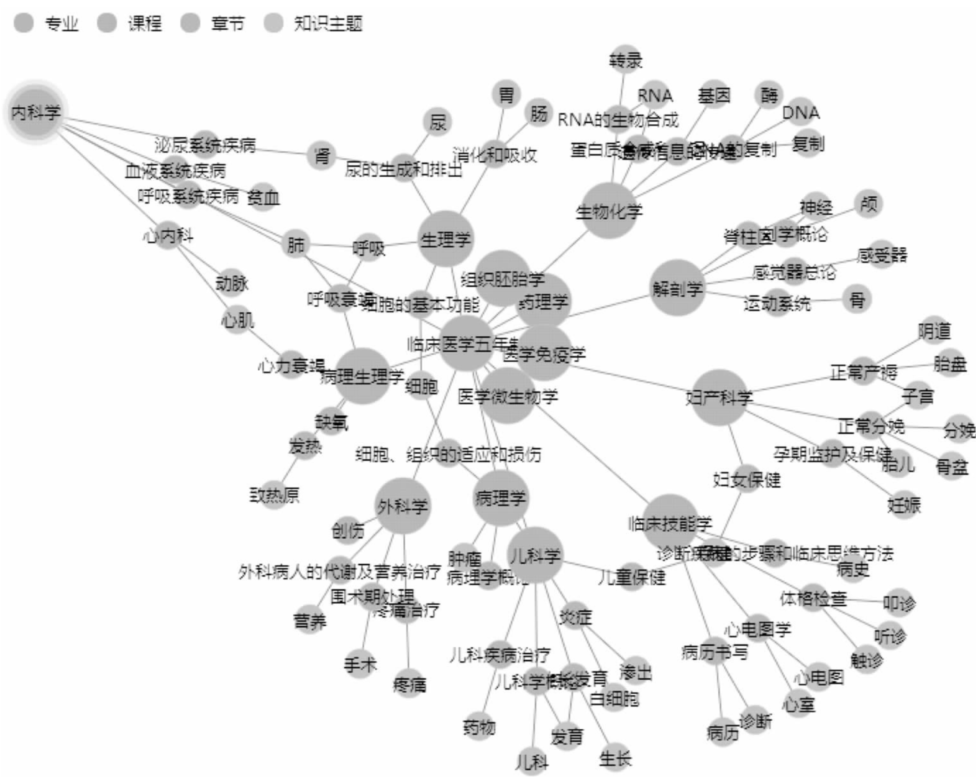
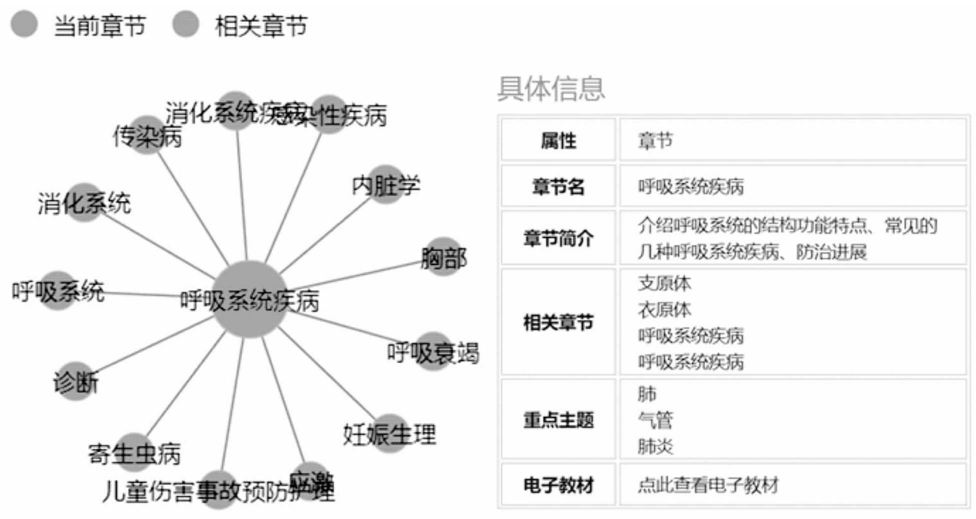


图 3　临床医学课程知识主题图谱总览

相关的章节名称。如图 4 所示，在《内科学》课程中的“呼吸系统疾病”图谱中与其关联较大的章节有“内脏学”“胸部”“呼吸衰竭”“呼吸系统”“应激”等，图中章节与中心点的距离表明章节间的关联程度，距离越近则表示与“呼吸系统疾病”章节关联程度越大。具体信息中呈现该章节的章节简介、相关章节、重点主题等。



属性	章节
章节名	呼吸系统疾病
章节简介	介绍呼吸系统的结构功能特点、常见的几种呼吸系统疾病、防治进展
相关章节	支原体 衣原体 呼吸系统疾病 呼吸系统疾病
重点主题	肺 气管 肺炎
电子教材	点此查看电子教材

图 4　“呼吸系统疾病”相关章节及章节信息

3.3　主题—主题关联

对于一个主题节点，图谱呈现与该主题相关的主题名称。如图 5 所示，对于“哮喘”主题，其基础主题有“胸廓”“急性上呼吸道感染”“鼻炎”“支气管炎”等，表明学习该主题之前需要具备基础主题词的知识；进阶主题为“肺炎”，肺炎可能会引发哮喘，两者之间既有区别又有联系，在学习“哮喘”后应继续了解“肺炎”相关知识；同级主题为“胆碱”“衣原体”和“支原体”，则“哮喘”与其不存在偏序关系，可以同步学习。具体信息中呈现该主题的主题简介和所属章节等，儿科学第 12 章、内科学第 2 章和药理学第 5 章都包含了该主题。

陆泉，谢祎玉，陈静，等. 临床医学课程知识主题图谱构建研究[J]. 图书情报工作,2019,63(9):101－108.

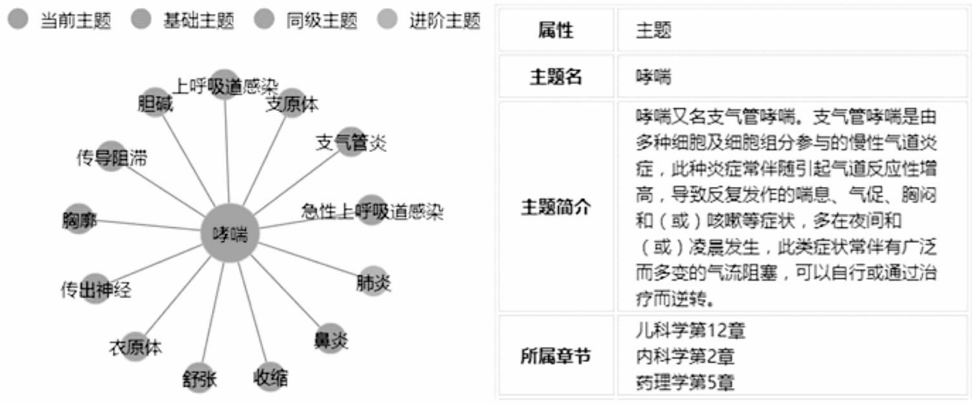


图 5 “哮喘”相关主题及主题信息

4 结语

本研究通过对武汉大学临床医学五年制专业主干课程的课程资料进行深度挖掘，得到临床医学课程间、知识主题间及课程与知识主题间的关联，揭示课程之间细粒度知识主题关联，构建了临床医学课程知识主题知识库及临床医学课程知识主题图谱，并在此基础上开发查询系统进行知识图谱可视化呈现，系统可以通过网址 http://218.197.150.149/rainbow 进行访问。

在理论层面，本研究将文本挖掘技术与情报学相关理论结合到专业课程知识体系研究中，从专业课程体系与知识主题的角度来构建特定领域的知识体系，是现有知识图谱理论的有益补充。在应用层面，深入挖掘临床医学课程知识主题图谱，可以打通专业内多课程间的知识壁垒，有助于教学管理人员及师生掌握专业知识体系，开展知识导向型教学活动。一方面，本研究成果可以为教学管理人员科学管理专业知识体系、系统优化课程体系、辅助教学排课与教学团队建设等提供关键的理论与技术基础；另一方面，可以将其应用于实际教学中，对于辅助教师合理组织专业知识点、优化教学计划、提高教学质量，学生深入理解课程之间细粒度知识主题关联、合理规划和系统学习，促进广大师生对学科知识的理解、利用与升华，对于我国专业人才培养、医学领域知识组织与服务及智慧医学教育等方面具有普遍意义与应用价值。

本研究仍存在诸多不足之处，如在计算章节—章节关联时，仅考虑知识主题词共现的关联规则，没有考虑其在两个章节中的分布特征，可能无法准确揭示章节之间的关联权重，因此可以结合关联规则和主题词分布特征对算法进一步改进；另外，已有的课程体系可能是不准确的，可以结合专业领域知识图谱（如 Linked Life Data 上的子图分析）来对现有的图谱进行优化。

参考文献：

[1] 杨洋，樊玉霞，王丹，等. 临床医学五年制本科课程改革中有关系统化教学的探讨[J]. 试题与研究：教学论坛，2016(22)：19－19.

[2] 教育部，卫生部. 本科医学教育标准——临床医学专业（试行）[EB/OL]. [2018－07－24]. http://www.moe.gov.cn/srcsite/A08/moe_740/s3864/200809/t20080916_109605.html.

[3] 秦磊. 课程建设是内涵发展的重要抓手[N]. 中国教育报，2015－06－02(10).

[4] 胡文韬. 基于知识图谱的学习路径图生成技术研究[D]. 北京：北京邮电大学，2017.

[5] 叶春森，汪传雷，储节旺. 基于知识地图的知识集成模式与机理研究[J]. 情报理论与实践，2009，32(10)：52－54.

[6] 郑宁. 基于自然语言处理的程序设计资源解题知识发现研究[D]. 上海：东华大学，2014.

[7] 商玮，盘红华，郭飞鹏. TF-CDIO 电子商务专业课程体系的构建[J]. 高等工程教育研究，2012(2)：146－151.

[8] 周明，谢俊. 大数据视角下信管专业的培养模式创新研究[J]. 图书馆学研究，2016(6)：41－46.

[9] 李莉，白大勇，张诚玥，等. 基于大数据技术的眼科教学体系建设探讨[J]. 中国医院管理，2015，35(8)：51－53.

[10] TODD R J. From information to knowledge：charting and measuring changes in students' knowledge of a curriculum topic[J]. Information research an international electronic journal，2006，11(4)：264－264.

[11] 朱珂，刘清堂，叶阳梅. 基于主题图的网络课程知识组织研究[J]. 电化教育研究，2014(1)：91－96.

[12] MELIS E，ANDRES E，FRISCHAUF A，et al. ActiveMath：a generic and adaptive web-based learning environment[J]. International journal of artificial intelligence in education，2001，12(4)：385－407.

[13] 武汉大学医学部. 临床医学（五年制）本科人才培养方案（2013 版）[EB/OL]. [2018－07－25]. http://wsm70.whu.edu.cn/content1.jsp? urltype = news.NewsContentUrl&wbtreeid =

1055&wbnewsid=7811.
[14] 中国医学科学院. 中国生物医学文献数据库[EB/OL]. [2018-11-23]. http://www.sinomed.ac.cn/zh/.
[15] BLEI D M, Ng A Y, JORDAN M I. Latent dirichlet allocation [J]. Machine learning research archive, 2003, 3:993-1022.
[16] HOFFMAN M D, BLEI D M, BACH F R. Online learning for latent Dirichlet allocation[C]// International conference on neural information processing systems. New York: Curran Associates Inc., 2010: 856-864.
[17] 唐晓波, 房小可. 基于文本聚类与LDA相融合的微博主题检索模型研究[J]. 情报理论与实践, 2013, 36(8):85-90.
[18] GRIFFITHS T L, STEYVERS M. Finding scientific topics[J]. Proceedings of the national academy of sciences of the United States of America, 2004, 101(S1):5228-5235.
[19] 朱泽德, 李淼, 张健, 等. 一种基于LDA模型的关键词抽取方法[J]. 中南大学学报(自然科学版), 2015(6):2142-2148.
[20] 唐晓波, 向坤. 基于LDA模型和微博热度的热点挖掘[J]. 图书情报工作, 2014, 58(5):58-63.
[21] 胡吉明, 陈果. 基于动态LDA主题模型的内容主题挖掘与演化[J]. 图书情报工作, 2014, 58(2):138-142.
[22] GUO X, XIANG Y, CHEN Q, et al. LDA-based online topic detection using tensor factorization[J]. Journal of information science, 2013, 39(4):459-469.
[23] 史忠植. 知识发现[M]. 北京:清华大学出版社, 2002.
[24] 张晗, 路振宇, 崔雷. 利用关联规则对医学文本数据库进行知识抽取的尝试——以四种抗肿瘤药为例[J]. 现代图书情报技术, 2006, 1(9):49-52.
[25] HAN J. Data mining: concepts and techniques[M]. San Francisco: Morgan Kaufmann Publishers Inc., 2005.
[26] INOKUCHI A, WASHIO T, MOTODA H. An apriori-based algorithm for mining frequent substructures from graph data[C]//European conference on principles of data mining and knowledge discovery. Berlin: Springer-Verlag, 2000:13-23.
[27] 钟伟金, 李佳. 共词分析法研究(一)——共词分析的过程与方式[J]. 情报杂志, 2008, 27(5):70-72.
[28] 高继平, 丁堃, 潘云涛, 等. 多词共现分析方法的实现及其在研究热点识别中的应用[J]. 图书情报工作, 2014, 58(24):80-85.
[29] WU H C, LUK R W P, WONG K F, et al. Interpreting TF-IDF term weights as making relevance decisions[J]. ACM transactions on information systems, 2008, 26(3):55-59.
[30] 宋美娜, 崔丹阳, 鄂海红, 等. 一种通用的数据可视化模型设计与实现[J]. 计算机应用与软件, 2017, 34(9):38-42.

作者贡献说明:

陆泉:提出研究问题,设计研究框架,提供论文修改建议;
谢祎玉:进行数据处理,撰写论文主体和修改论文;
陈静:提出研究思路,修订论文;
张涵:进行模型实现和数据处理;
崔浩冉:进行知识图谱可视化;
聂书源:调研和采集研究所需数据。

Research on the Construction of Clinical Medicine Course-knowledge Topic Graph

Lu Quan[1,2] Xie Yiyu[1] Chen Jing[3] Zhang han[1] Cui Haoran[1] Nie Shuyuan[1]
[1] Center for Studies of Information Resources of Wuhan University, Wuhan 430072
[2] Big Data Institute, Wuhan University, Wuhan 430072
[3] School of Information Management, Central China Normal University, Wuhan 430079

Abstract: [**Purpose/significance**] From the perspective of knowledge topics, this paper try to solve the problems of overlaps and "information island" between courses in the existing curriculum system design and teaching by establishing a comprehensive curriculum knowledge system. Thus, the professional knowledge services can be carried out effectively. [**Method/process**] This study takes the main courses of clinical medicine as the research object, based on medical thesis vocabulary, electronic textbooks, electronic lesson plans and other medical education data, through the LDA model to deeply explore the knowledge topics in courses, and then using the correlation analysis method to reveal the fine-grained relationship between courses, knowledge topics and the courses and knowledge topics. Thus, a clinical medical course-knowledge topic graph is constructed. [**Result/conclusion**] The study constructs domain knowledge graph from the perspective of professional curriculum system and knowledge subject. The results will help teaching managers, teachers and students master the professional knowledge system, and carry out knowledge-oriented teaching activities. Furthermore, It can promote the development of knowledge organization and services in the medical field and the development of smart medical education.

Keywords: course-knowledge topic graph knowledge graph LDA correlation analysis clinical medicine